Data Management at NIST

John Henry J. Scott[1]
Raymond Plante
Office of Data and Informatics (ODI)
Material Measurement Laboratory
National Institute of Standards and Technology

Tuesday, March 20, 2018
Scientific Data Management for Photon and Neutron Facilities
HZB, Berlin

[1] johnhenry.scott@nist.gov
https://orcid.org/0000-0002-3796-2110

MATERIAL MEASUREMENT LABORATORY

NIST National Institute of Standards and Technology U.S. Department of Commerce

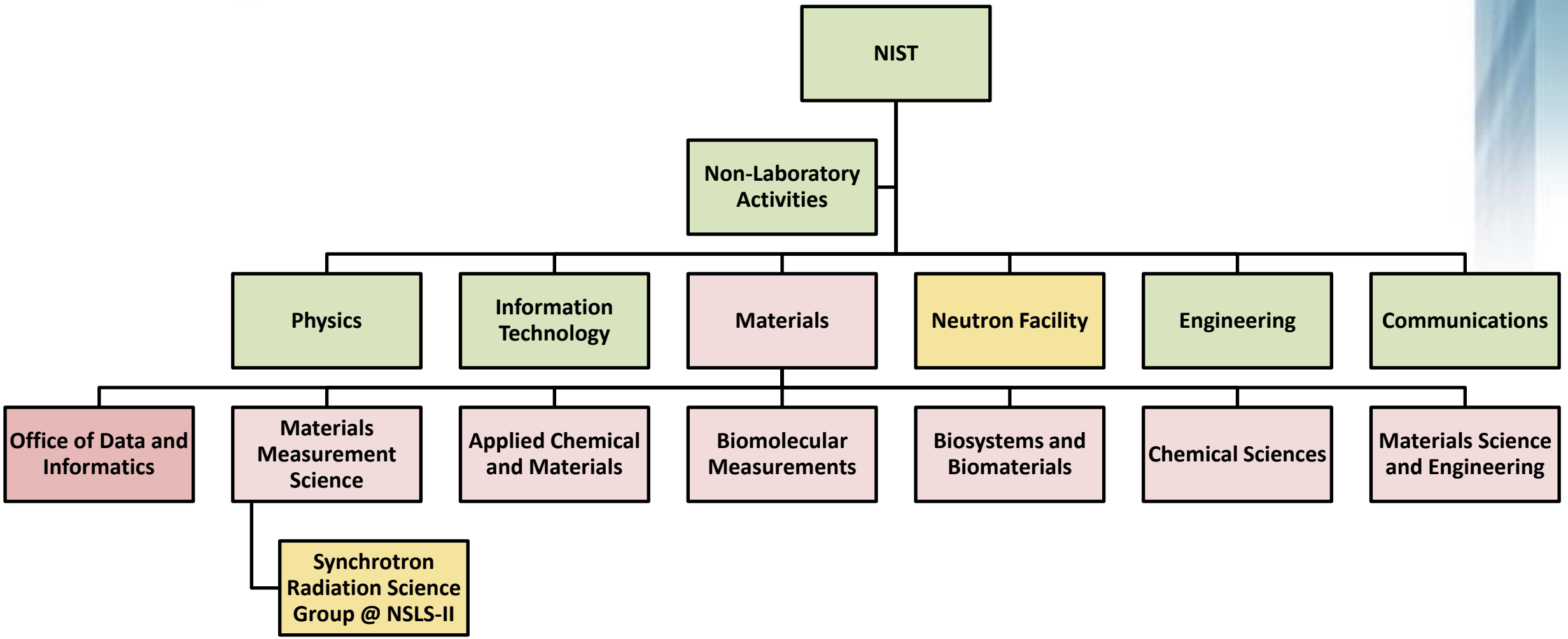
Credits

- ODI Staff and Details: Gretchen Greene, Adam Morey, Chandler Becker, Peter Linstrom, Arlin Stoltzfus, Kim Tryka, Andrea Medina-Smith, Zachary Trautt, June Lau, et al.
- ITL Staff: Mary Brady, Alden Dima, Sharief Youssef, et al.
- Others: Office of Information Systems Mgt, Kathy Sharpless, many others

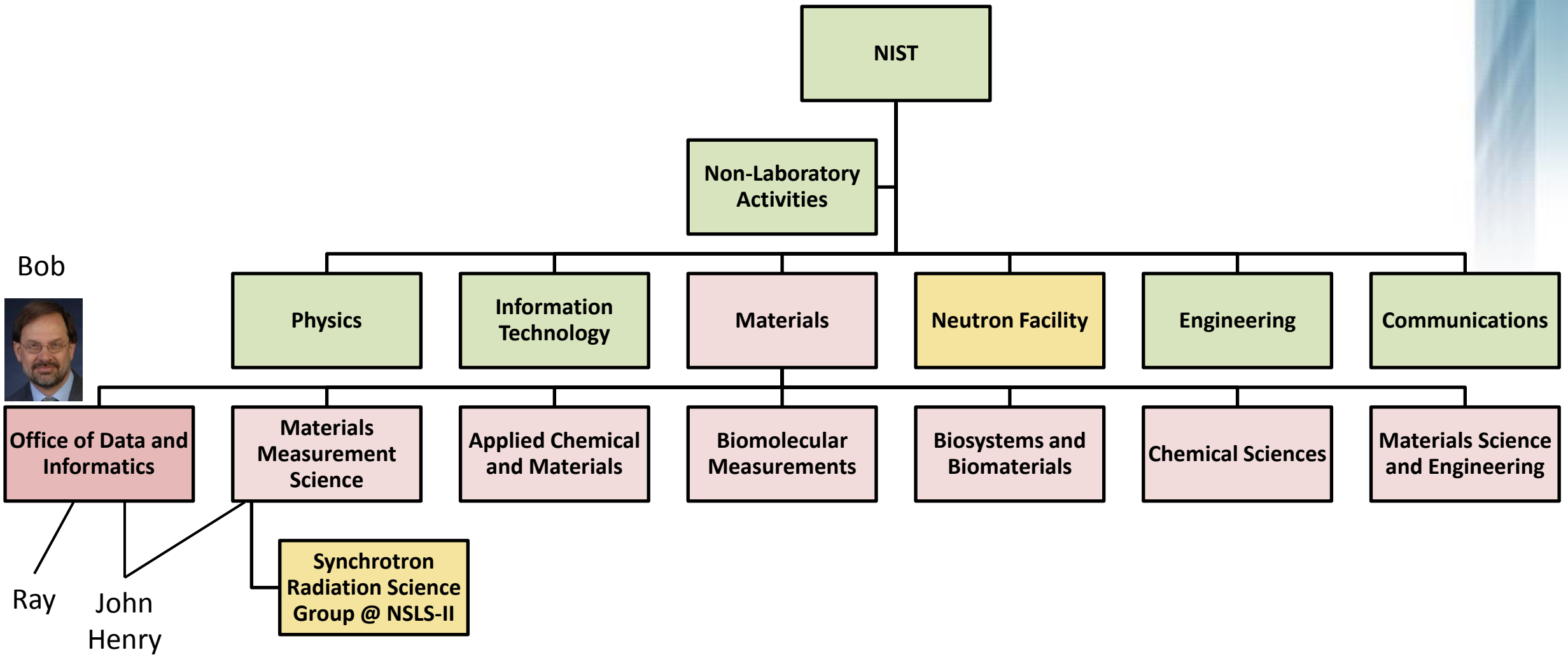
Challenges

- Incredible diversity of research programs
  - Size, complexity, discipline
  - Most data generated by small research groups (long tail)
- Conservative culture, resistant to change
  - Must demonstrate high ratio of benefit to cost
- Little recognition for data sharing/publication for advancement
- Proprietary data formats
  - Many bench instruments have opaque formats
  - Vendors want to sell proprietary software as well as the instrument
- IT security
  - Widely-used tools like Dropbox prohibited
  - New deployments require A&A, ATO – takes time and effort

NIST Organization



NIST Organization



MATERIAL MEASUREMENT LABORATORY

NIST National Institute of Standards and Technology U.S. Department of Commerce

Office of Data and Informatics

Standard Reference Data

- Distribution
- Sales
- Infrastructure
- Usage analysis and impact
- Improve web sites and user interfaces
- Provide APIs

Research Data

- Improve data management practices
- Data management planning tools
- Laboratory automation
- Electronic Lab Notebooks
- NIST open data repository
- NIST data portal

Data Science

- Informatics and analytics resource
- Liaison with NIST Information Technology Laboratory
- Big data
- Cloud computing
- National Strategic Computing Initiative

Community

- National Data Services Consortium
- Research Data Alliance
- Other US federal agencies (NIH, DOE, NSF)
- CENDI
- National Metrology Institutes (NMIs)
- BIPM (Paris)
- CODATA, WDS

Robert Hanisch
Director, Office of Data and Informatics

MATERIAL MEASUREME LABORATORY

NIST National Institute of Standards and Technology U.S. Department of Commerce

Data Management at NIST

Discover

- Materials Resource Registry (data, code)
- International Metrology Resource Registry
- NIST Enterprise Data Inventory
- data.gov
- NIST Public Data Repository and Search Portal

- Standard Reference Data
- Materials Data Repository
- Materials Data Facility
- Persistent identifiers (DOIs, handles)

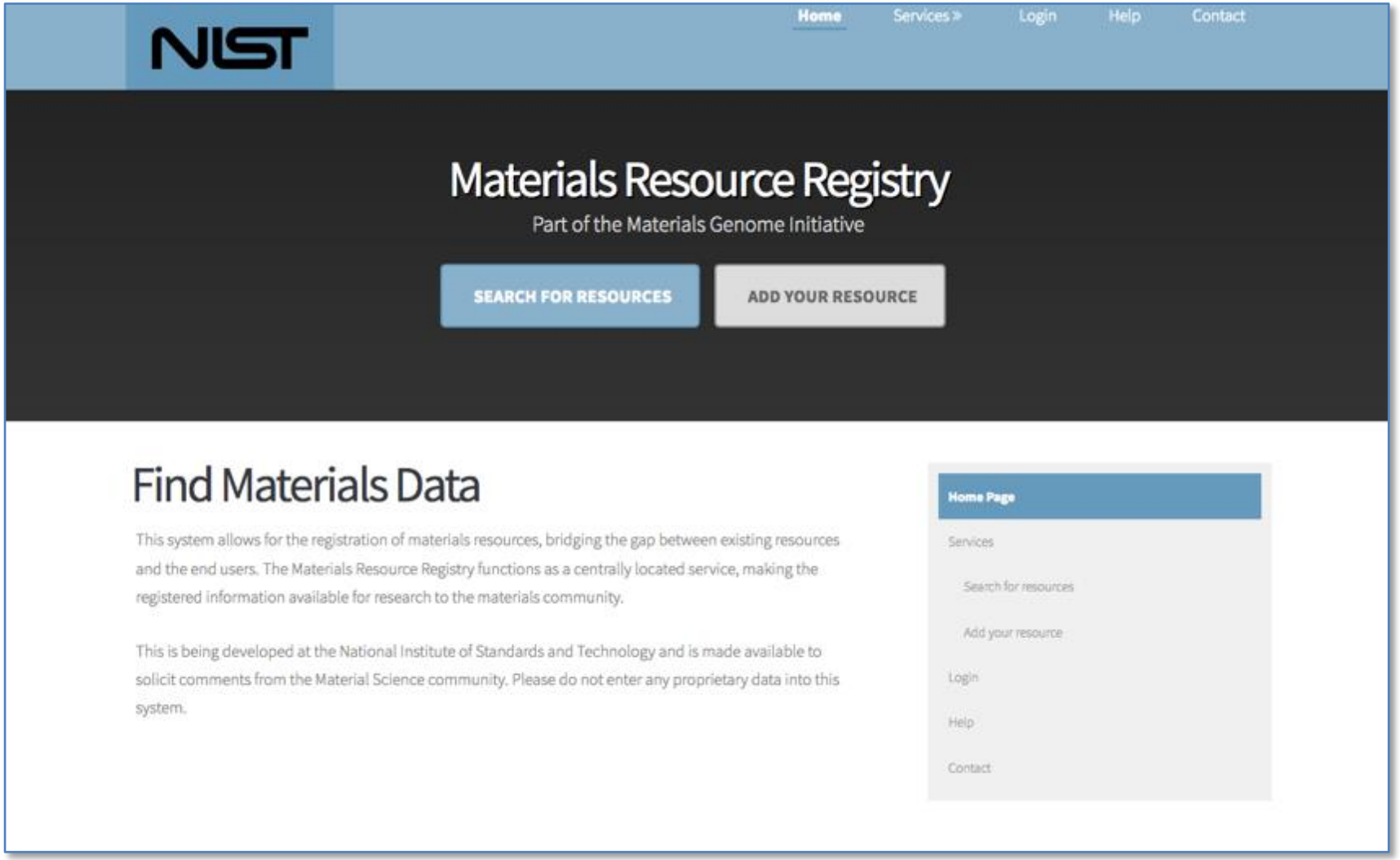
Access

Interoperate

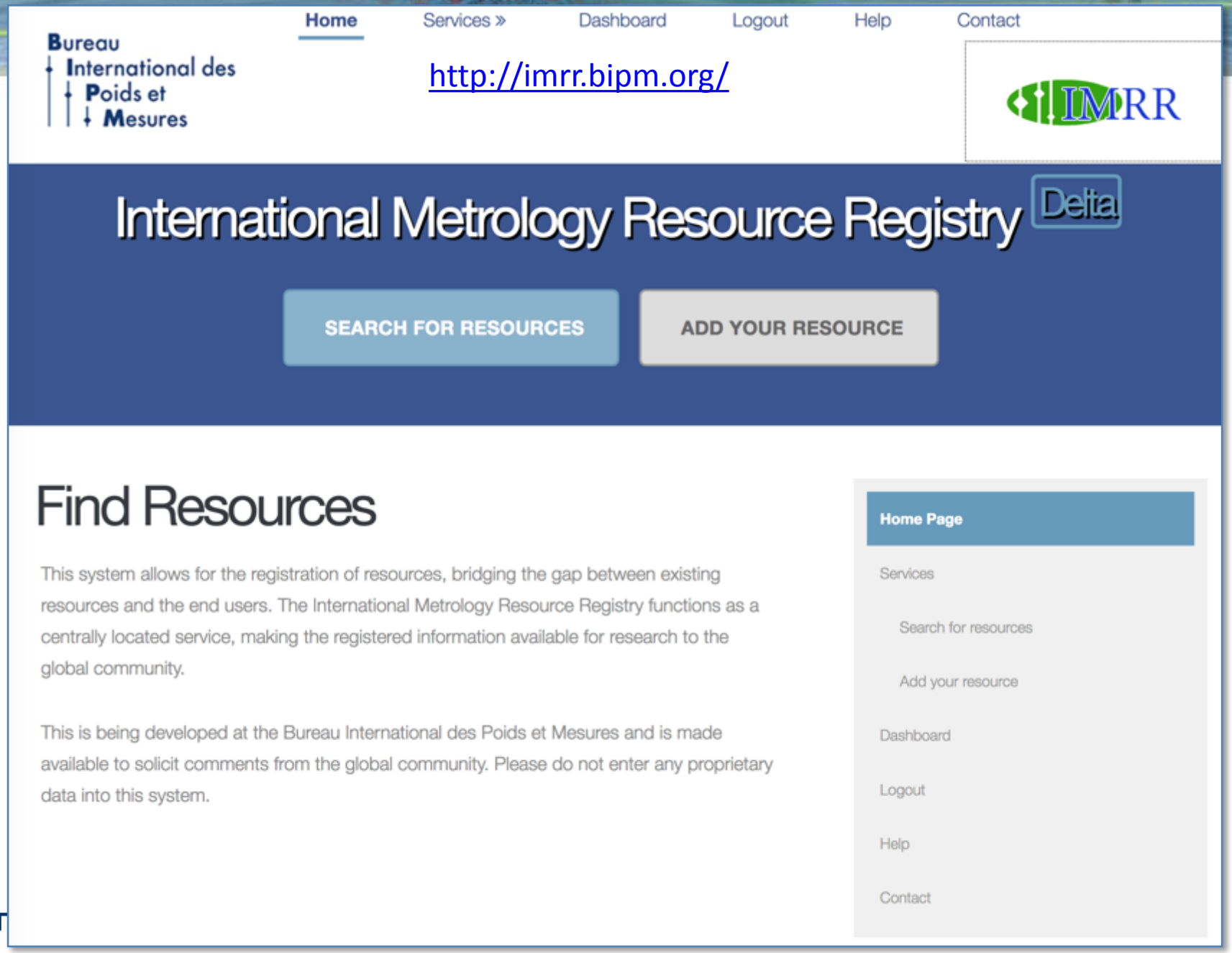
- Materials Data Curator
- Data type registry
- Schema repository
- Lab info mgmt systems

Discovery

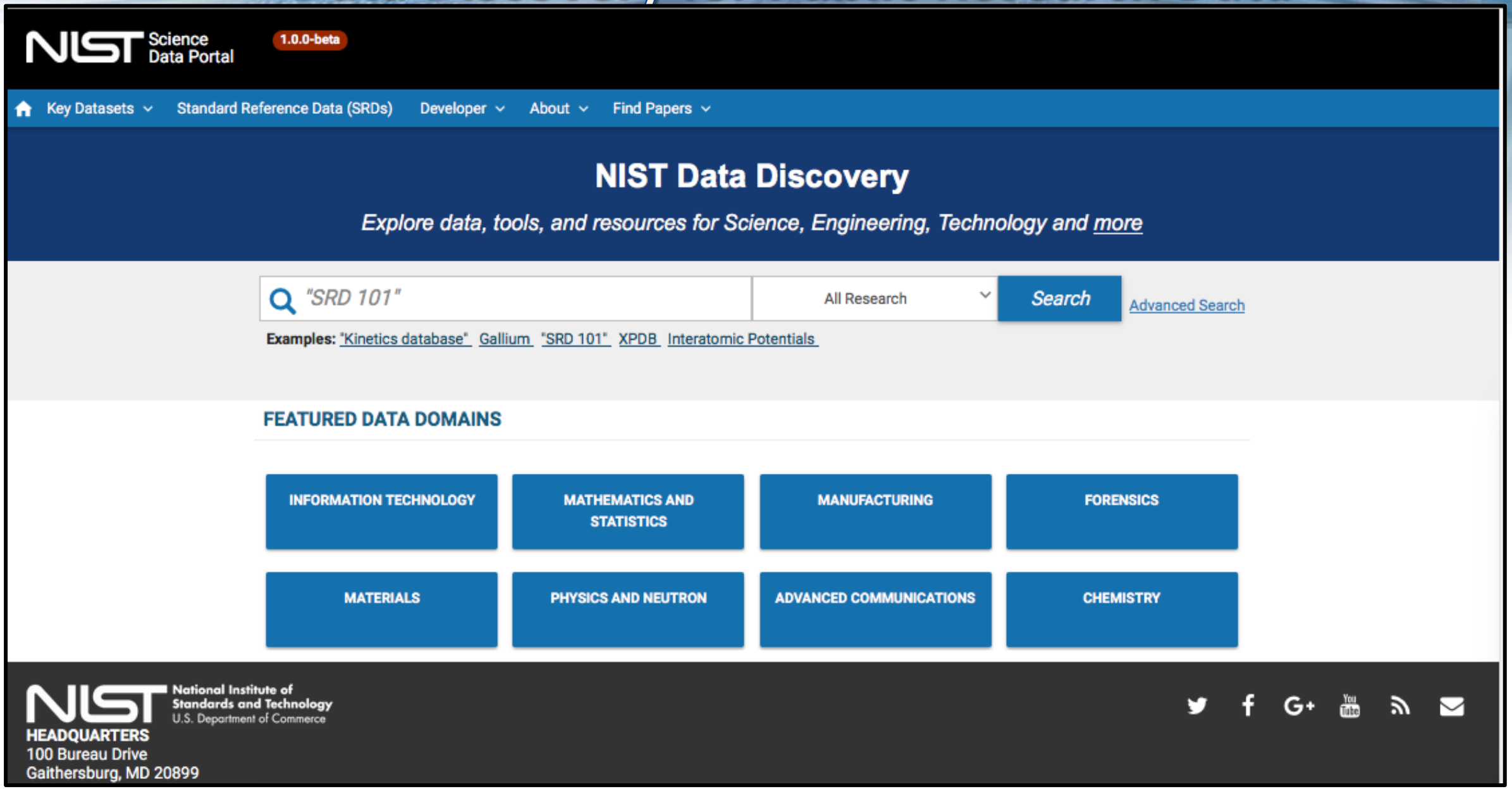
Materials Resource Registry



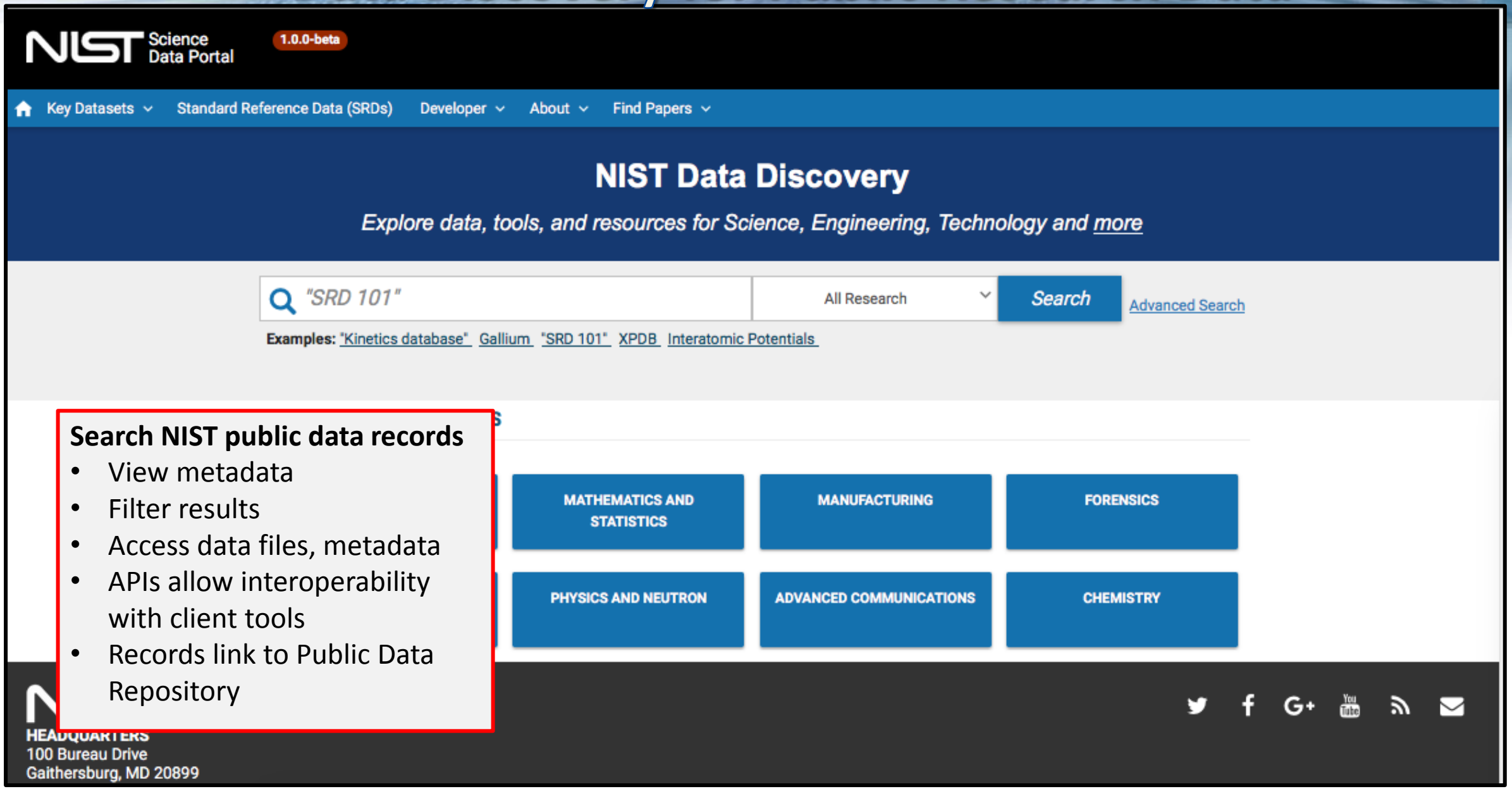
https://materials.registry.nist.gov/


Home
Services »
Dashboard
Logout
Help
Contact
Bureau International des Poids et Mesures
http://imrr.bipm.org/
IMRR
International Metrology Resource Registry
Delta
SEARCH FOR RESOURCES
ADD YOUR RESOURCE
Find Resources
This system allows for the registration of resources, bridging the gap between existing resources and the end users. The International Metrology Resource Registry functions as a centrally located service, making the registered information available for research to the global community.
This is being developed at the Bureau International des Poids et Mesures and is made available to solicit comments from the global community. Please do not enter any proprietary data into this system.
Home Page
Services
Search for resources
Add your resource
Dashboard
Logout
Help
Contact

Data Discovery for Public Research Data

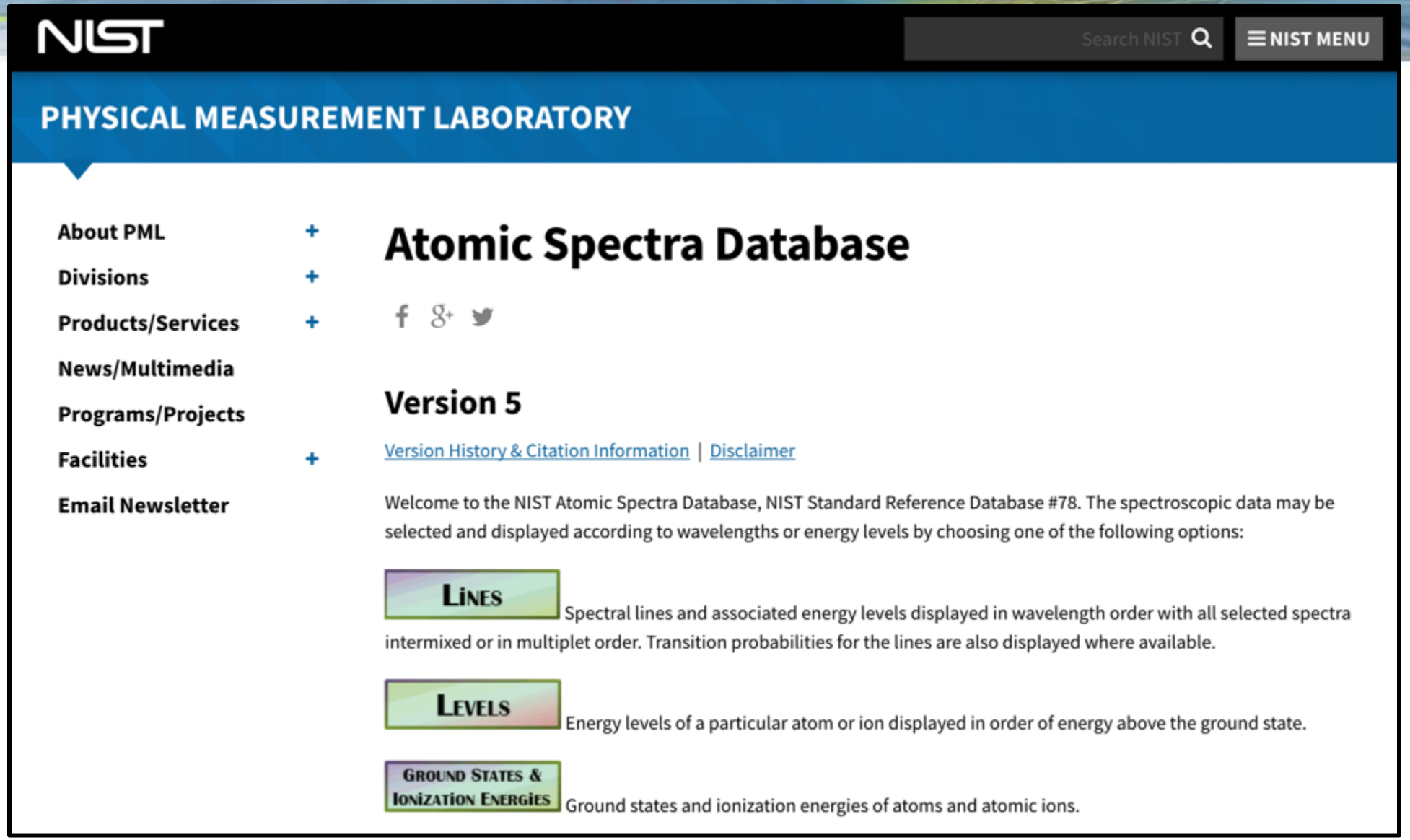


Data Discovery for Public Research Data

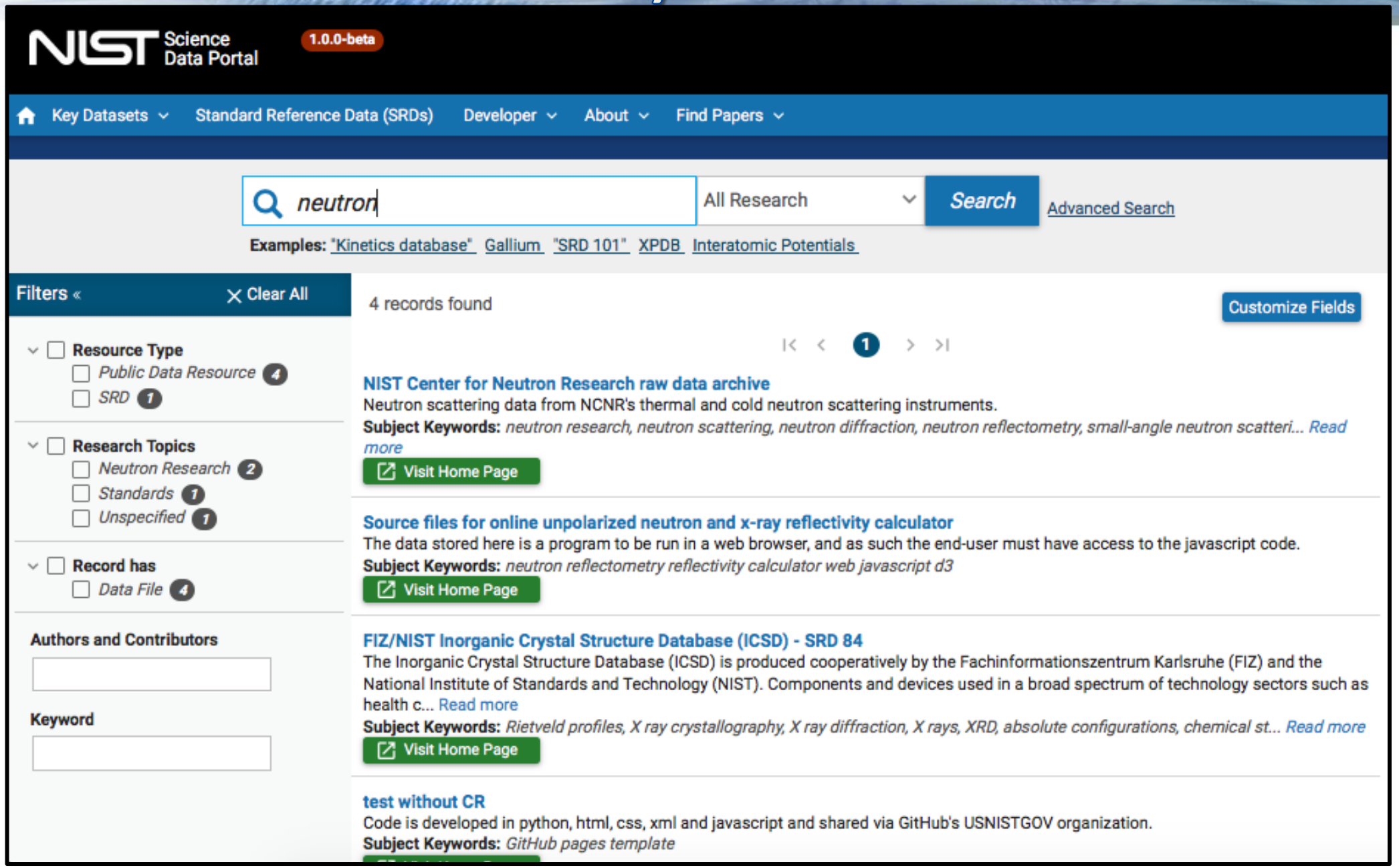


Search NIST public data records

- View metadata
- Filter results
- Access data files, metadata
- APIs allow interoperability with client tools
- Records link to Public Data Repository


NIST
Search NIST
NIST MENU
PHYSICAL MEASUREMENT LABORATORY
About PML +
Divisions +
Products/Services +
News/Multimedia
Programs/Projects
Facilities +
Email Newsletter
Atomic Spectra Database
Version 5
Version History & Citation Information | Disclaimer
Welcome to the NIST Atomic Spectra Database, NIST Standard Reference Database #78. The spectroscopic data may be selected and displayed according to wavelengths or energy levels by choosing one of the following options:
LINES
Spectral lines and associated energy levels displayed in wavelength order with all selected spectra intermixed or in multiplet order. Transition probabilities for the lines are also displayed where available.
LEVELS
Energy levels of a particular atom or ion displayed in order of energy above the ground state.
GROUND STATES & IONIZATION ENERGIES
Ground states and ionization energies of atoms and atomic ions.

Data Discovery for Public Research Data



NIST Public Data Repository – Basic Landing Page

NIST Public Data Repository 1.0.0-beta

About | Search

Public Data Resource

FIB SEM image data set of Caenorhabditis elegans exposed to 60 nm Au nanoparticles

Contact: Keana Scott ..
Identifier: doi:10.18434/M3C09F
Last modified: **2015-09-20**

Description

This folder contains image data sets from 14 separate serial sectioning sessions. The entire data folder consists of 1379 8 bit tif images and is 47.3 GB in size. Serial sectioning was performed using FEI Helios 660 NanoLab focused ion beam scanning electron microscope (FIB SEM) and Auto Slice and View G3 software. The sample was a heavy metal stained and resin embedded Caenorhabditis elegans (C. elegans) that were exposed to 60 nm Au nanoparticles. Detailed descriptions of the worm preparation and resin block processing are described in Johnson, M.E. et al. (ACS Nano, 2016). Although the images were collected over 14 different sessions, they represent a contiguous section of a worm.

Subject Keywords: Advanced Materials, Biosciences and Health, Environment and Climate, Manufacturing, Visualization Research, Nanotechnology

References:

This data is referenced in :
http://pubs.acs.org/doi/abs/10.1021/acsnano.6b06582

Access To Data:

This data is public.

Access
- Visit Home Page
- Download all data
- Export JSON

Use
- Cite this resource
- License Statement

Find
- Similar Resources
- Resources by Authors

Table of Contents
- Description
- Files
- Metadata

NIST Public Data Repository – Basic Landing Page

NIST Public Data Repository 1.0.0-beta

About | Search

Public Data Resource

FIB SEM image data set of Caenorhabditis elegans exposed to 60 nm Au nanoparticles

Contact: Keana Scott ..
Identifier: doi:10.18434/M3C09F
Last modified: **2015-09-20**

Description

This folder contains image data sets from 14 separate serial sectioning sessions. The entire data folder consists of 1379 8 bit tif images and is 47.3 GB in size. Serial sectioning was performed using FEI Helios 660 NanoLab focused ion beam scanning electron microscope (FIB SEM) and Auto Slice and View G3 software. The sample was a heavy metal stained and resin embedded Caenorhabditis elegans (C. elegans) that were exposed to 60 nm Au nanoparticles. Detailed descriptions of the worm preparation and resin block processing are described in Johnson, M.E. et al. (ACS Nano, 2016). Although the images were collected over 14 different sessions, they represent a contiguous section of a worm.

Subject Keywords: Advanced Materials, Biosciences and Health, Environment and Climate, Manufacturing, Visualization Research, Nanotechnology

References:

This data is referenced in :
http://pubs.acs.org/doi/abs/10.1021/acsnano.6b06582

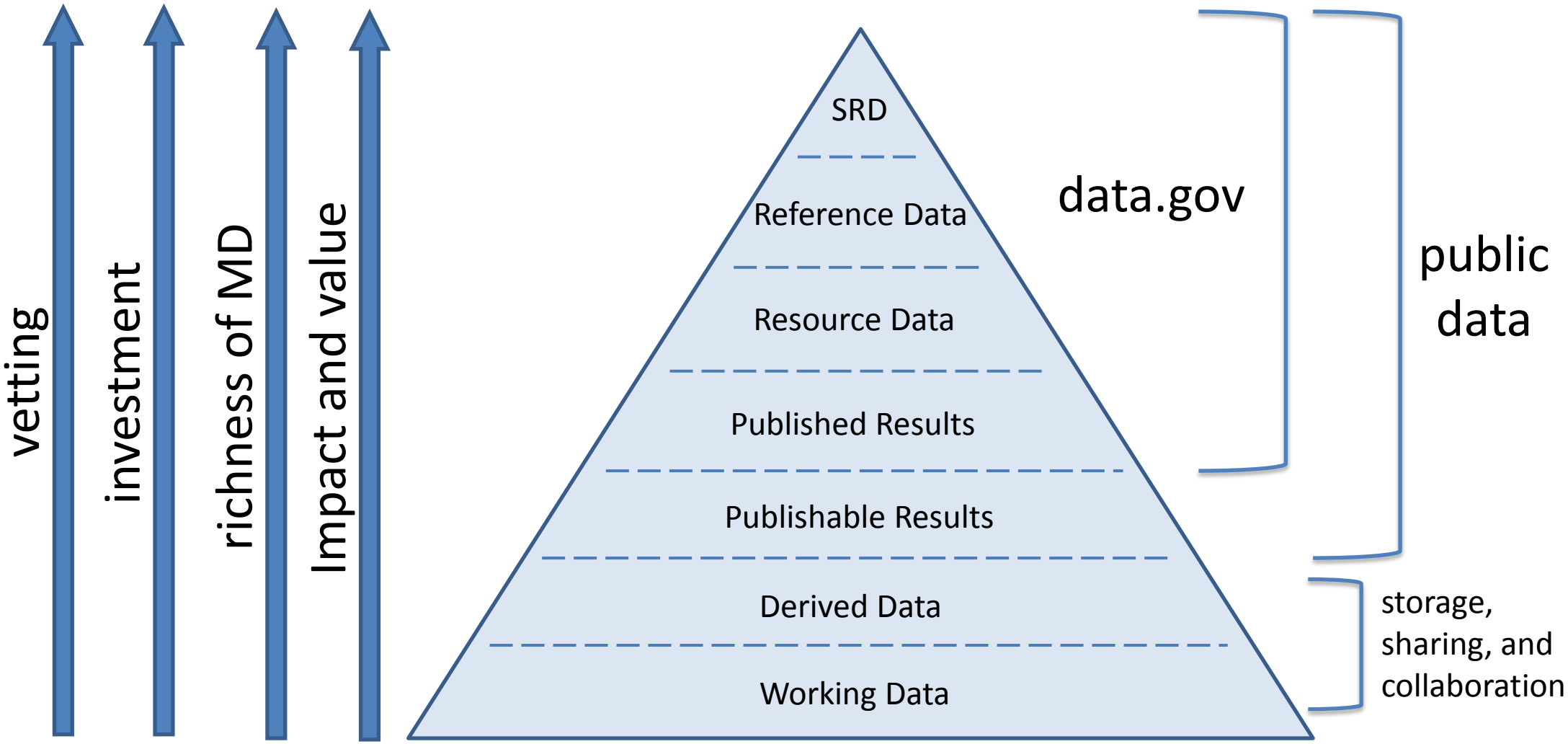
Access To Data:

This data is public.

- Published data landing page example
- The Public Data Repository provides capabilities for rich Data Publications
- Supports **NIST Extended metadata model** for domain science
- Standards for data publication are implemented

Access

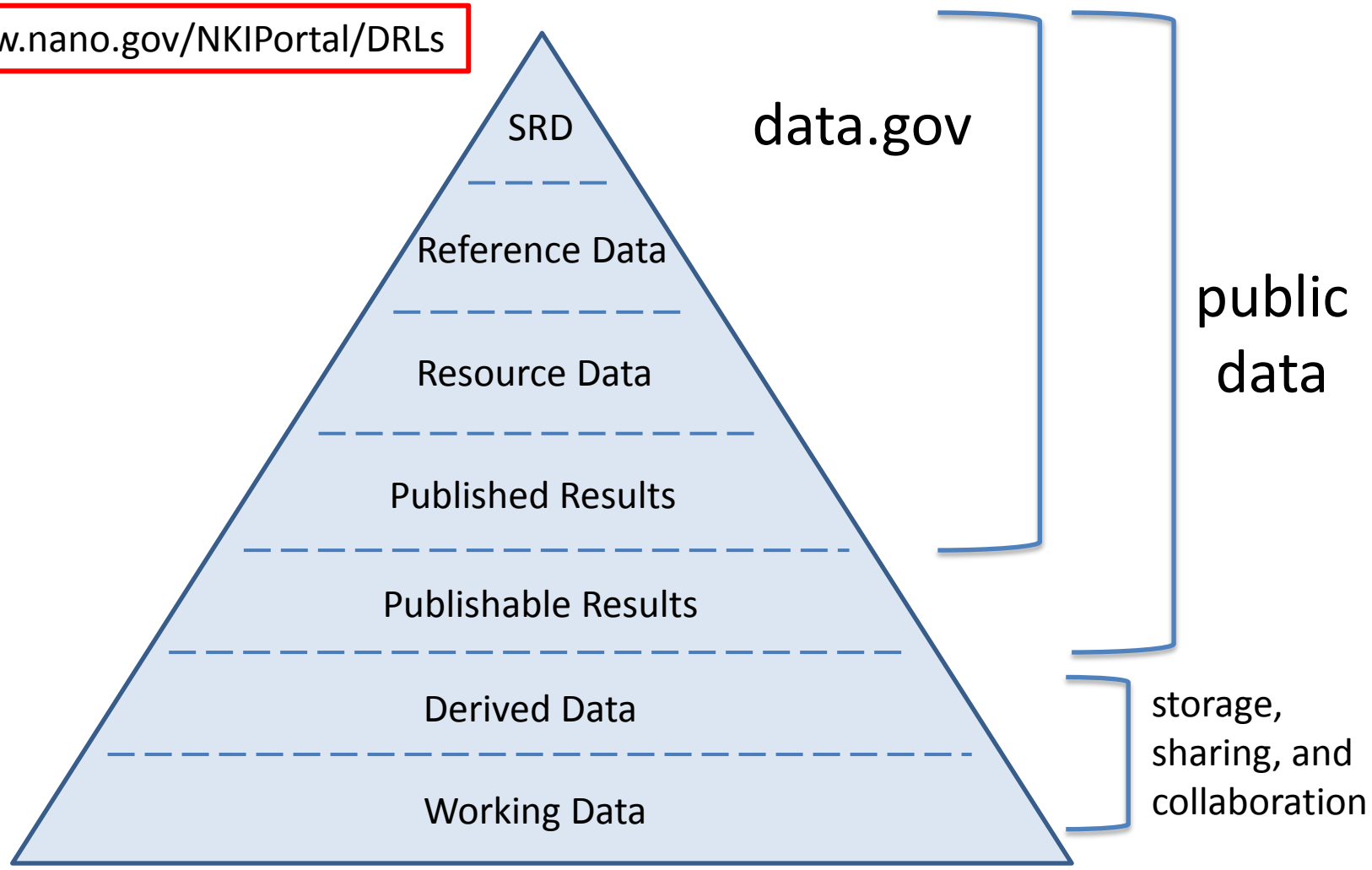
NIST Public Data Access Policy



NIST Public Data Access Policy

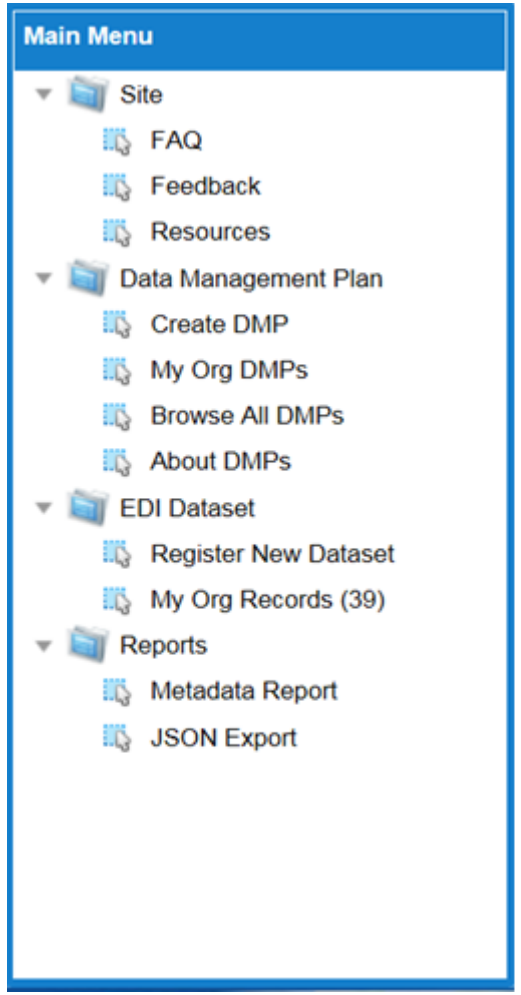
Data Readiness Level (DRL)
0. Invalid data
1. Raw or unscaled data
2. Scaled data
3. Scaled data with defined precision **or** noise level
4. Scaled data with defined precision **and** noise levels, but **not related** to the larger body of scientific knowledge
5. DRL 4 data **related** to the larger body of scientific knowledge, but with measurement uncertainty too large for data standards
6(X). Standards-quality data of X % measurement uncertainty

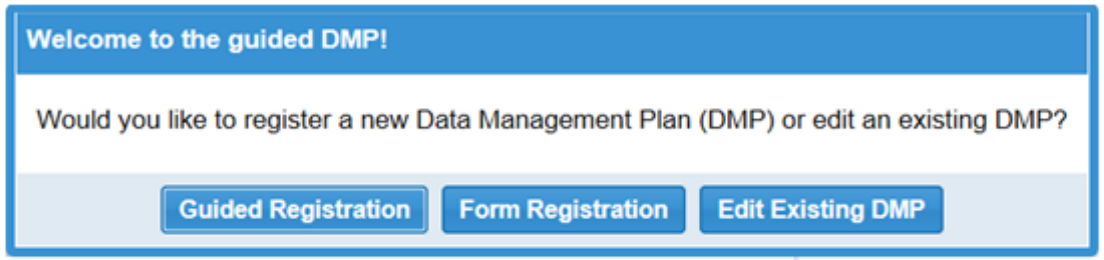
https://www.nano.gov/NKIPortal/DRLs

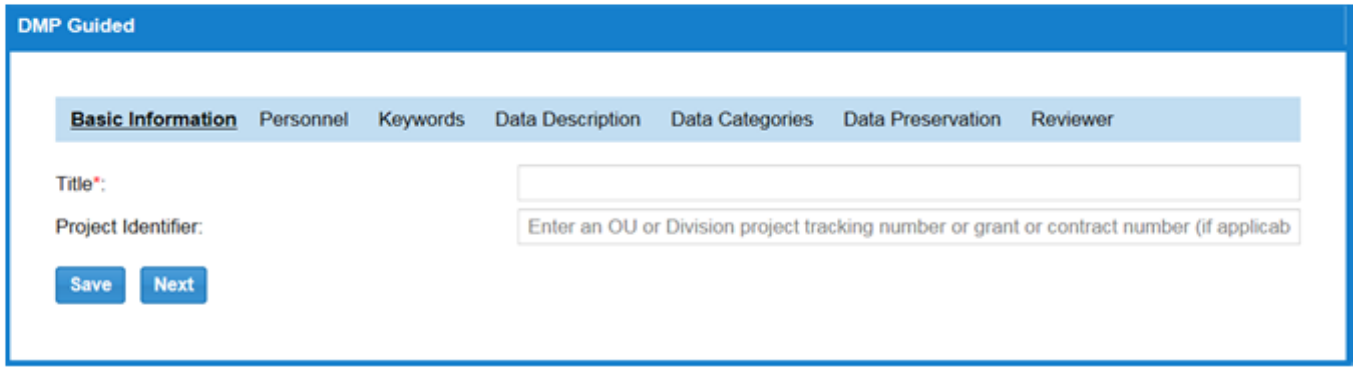


NIST Management of Institutional Data Assets (MIDAS)

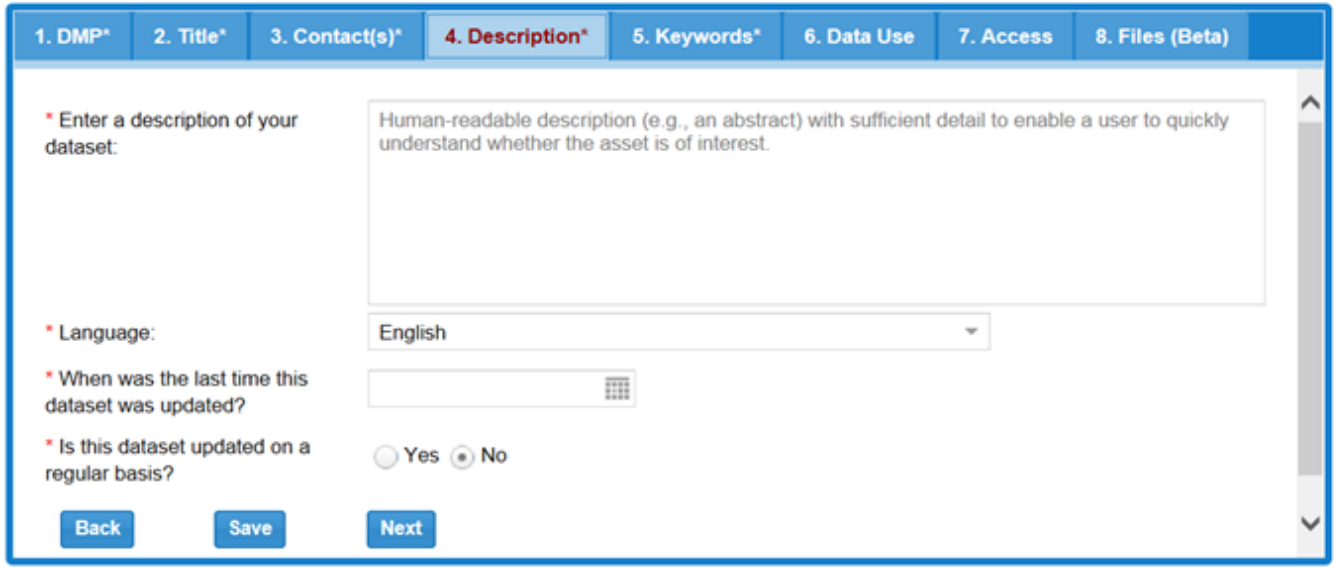
Creating a **Data Management Plan**



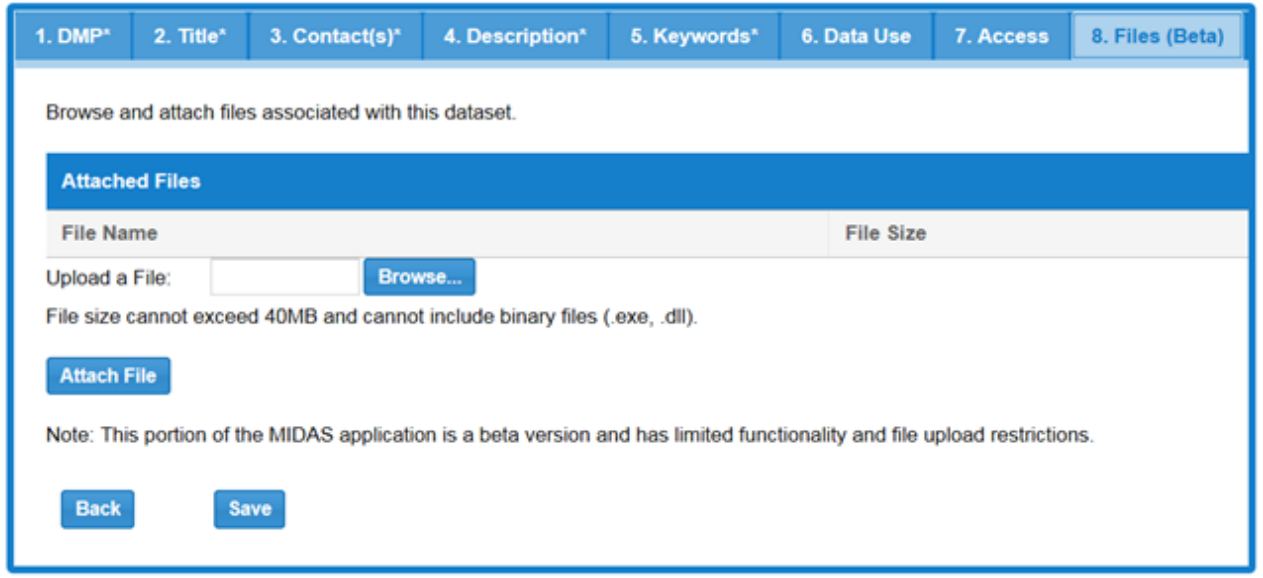




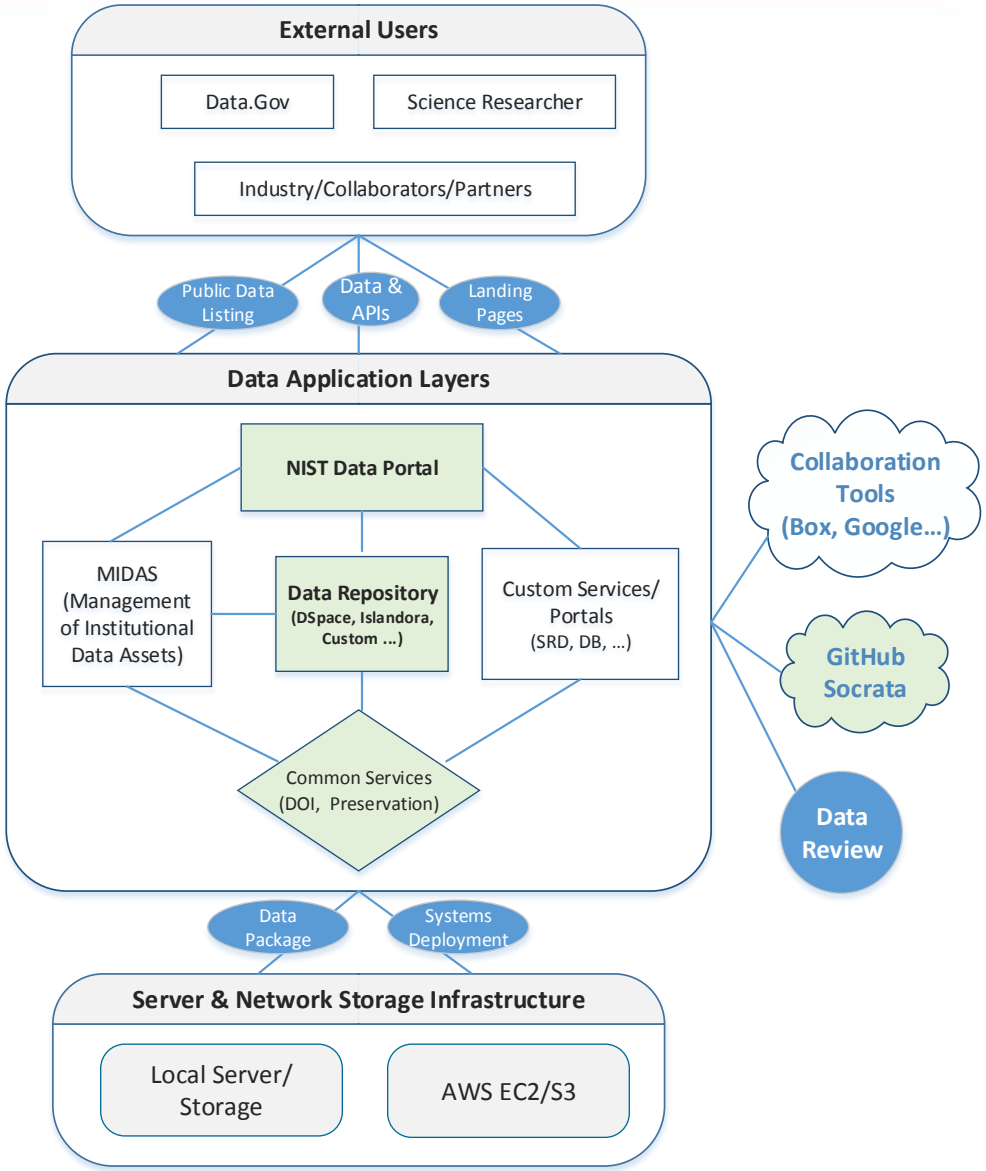
NIST Management of Institutional Data Assets (MIDAS)



Creating an **Enterprise Data Inventory** Record

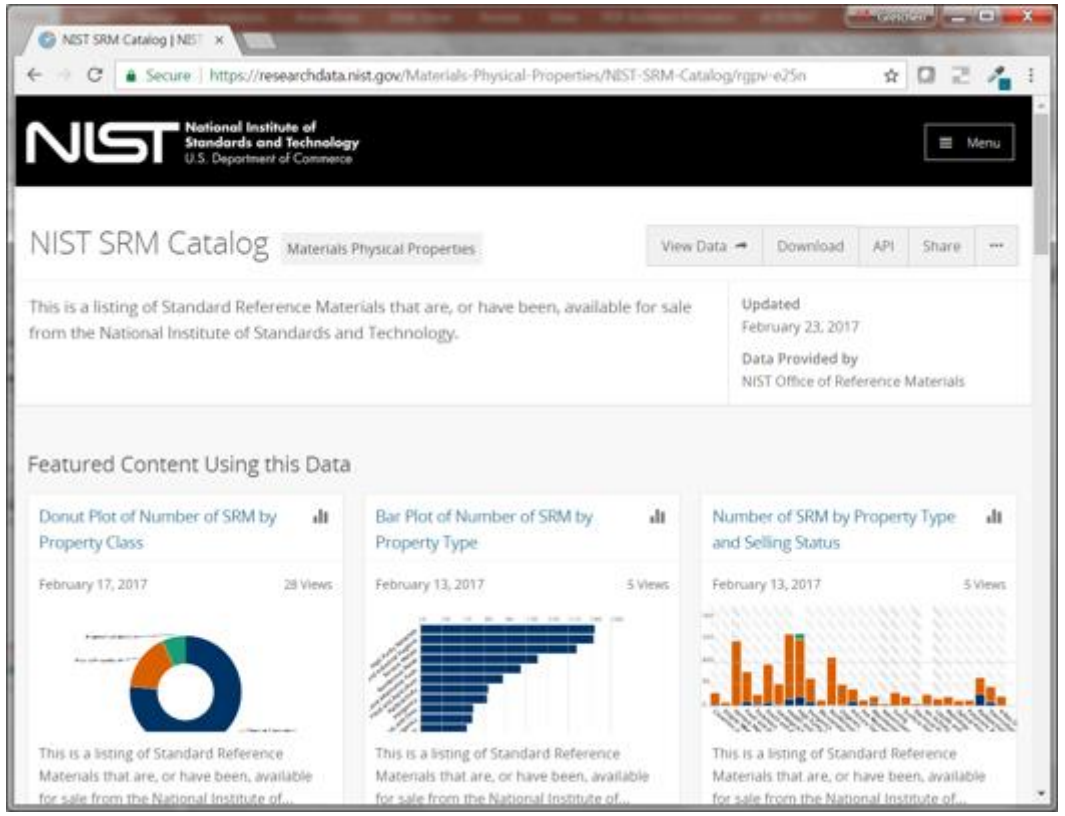


Research Data Infrastructure

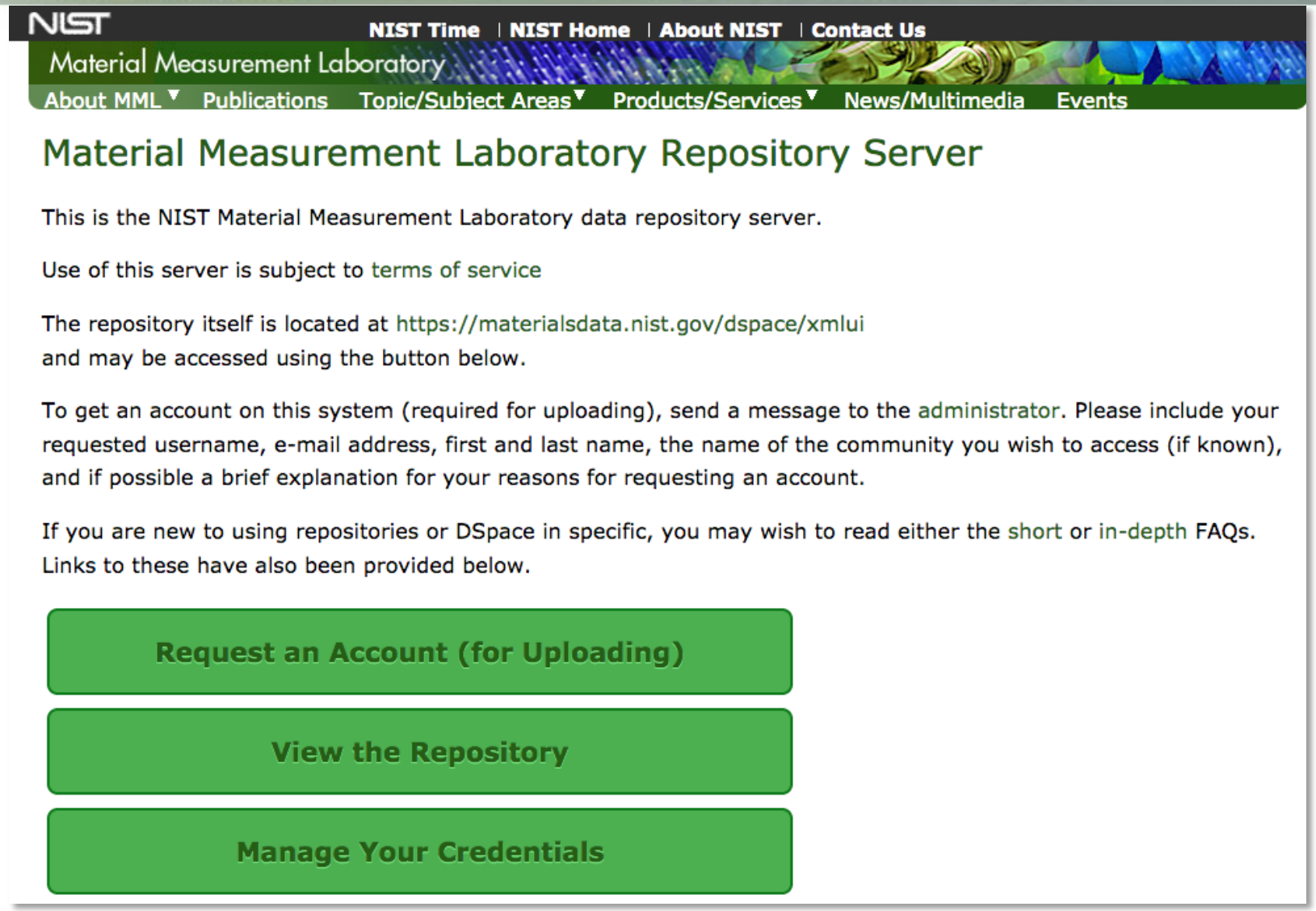


Sharing with Socrata Publishing Tool

- NIST public data records hosted in external Socrata publishing system
- Provides web landing page view and catalog for tabular datasets
- Search, filtering, and visualization tools
- Auto-generated API allows users to access data by script rather than by website



materialsdata.nist.gov



NIST Repositories

Communities in NIST Repositories

Select a community to browse its collections. [R] indicates an invitational community, [Z] indicates an archived community.

- ASM Structural Materials Data Demonstration Project
- Chemical Sciences
- CHiMaD Data Collections [R]
- Community for Greenhouse Gases
- Computational File Repository
- Experimental Data Repository
- Genome in a Bottle
- Heusler Phases: First Principles Simulations [R]
- ICME Approach to Development of Lightweight 3GAHSS Vehicle Assembly [R]
- ICME of Carbon Fiber Composites for Lightweight Vehicles [R]
- MGI Catalogs
- MICCoM Collections [R]
- NanoRelease [R]
- NIST/DOE-EERE Advanced Automotive Cast Magnesium Alloys [R]
- NIST Thermodynamics and Kinetics Test Space [R]
- Porous Metals and Ceramics: Freeze-casting under microgravity and terrestrial conditions
- RDA Demonstration Project: DTR/PID & MGI Infrastructure [R]
- RVE fracture VUMAT for QP980 steel
- State Variable Model for QP980
- Synchrotron Studies of Slot Die Coated Films [R]
- Thermal Conductivity of CVD Diamond - DARPA Round Robin [R]
- TMS Springer Integrating Materials and Manufacturing Innovation (IMMI)

Search NIST Repositories

Go

Advanced Search

Browse

All of NIST Repositories
- Communities & Collections
- By Issue Date
- Authors
- Titles
- Subjects

My Account

Login

Discover

Author
- Du, Y. (10)
- Abu-Farha, Fadi (8)
- Bower, Allan F. (8)
- Burton, Benjamin P. (7)
- van de Walle, Axel (7)
- Xu, Honghui (7)
- Zhang, L. (7)
- Liu, Shuhong (6)
- DeSchepper, Daniel C. (5)
- Flanagan, David P. (5)
- ... View More

Subject

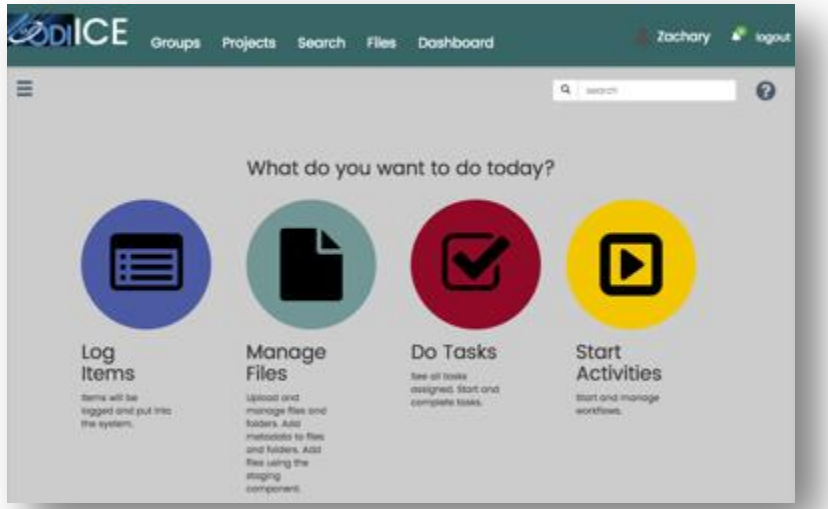
Interoperability

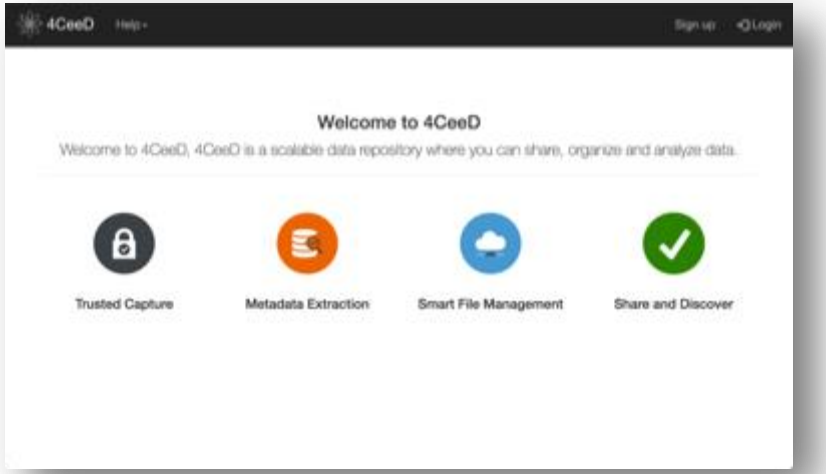
Laboratory Information Management Systems

- Integrated Collaborative Environment (ICE), a.k.a. Hyperthought
  - Running now at http://ice.nist.gov
  - Developed by Air Force Research Laboratory

- Timely and Trustworthy Curating and Coordinating Data Framework (T2C2) *4CeeD* system
  - Running now at http://t2c2.nist.gov:32500/
  - Developed by University of Illinois at Urbana-Champaign

- ~~Also considering Discovery Environment for Relational Information and Versioned Assets (DERIVA) from USC~~





Laboratory Information Management Systems

- Capture instrument metadata at the source
  - Metadata extractors
  - Often must reverse engineer proprietary binary formats
- Move experiment metadata into database
  - Enable search across many experiments
  - Do not use filenames/file system for metadata storage
- Enable scripted data processing, calibration, feature extraction
- Support data management from acquisition to publication; improve reproducibility

Materials Data Curation System

Materials Data Curation System

Part of the Materials Genome Initiative

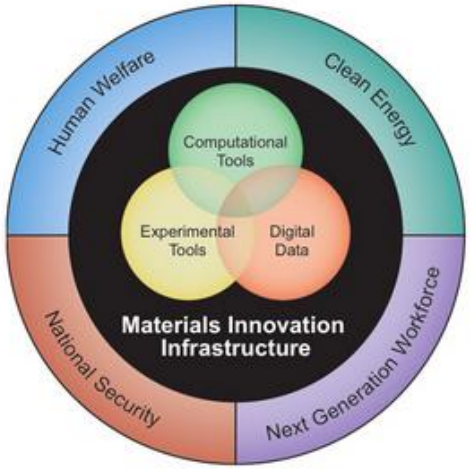
Login | My Profile | Help

Home

Materials Data Curator

This system allows for the curation of Material Data in a repository using predefined templates.

This is being developed at the National Institute of Standards and Technology and is made available to solicit comments from the Material Science community. Please do not enter any proprietary data into this system.



Available Options

All Options »

Most Recent Templates

Browse All »

species-qs | species.xsd.xml

TEM-Tutorial-Tamu | workshop-TEM.xsd

TEM-Tut | workshop-TEM.xsd

Materials Data Curation System

Digital Data
& Metadata
(any format)

Data Analysis
Infrastructure

GUI

Web
Framework

Exporter

REST API

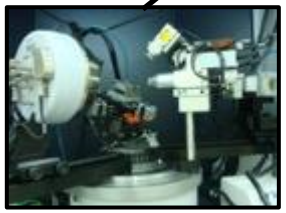
Data Provider

Harvester

Data Management
& Search Engine

Simulation

User Scripts



Measurement

Harvester

Data
Metadata

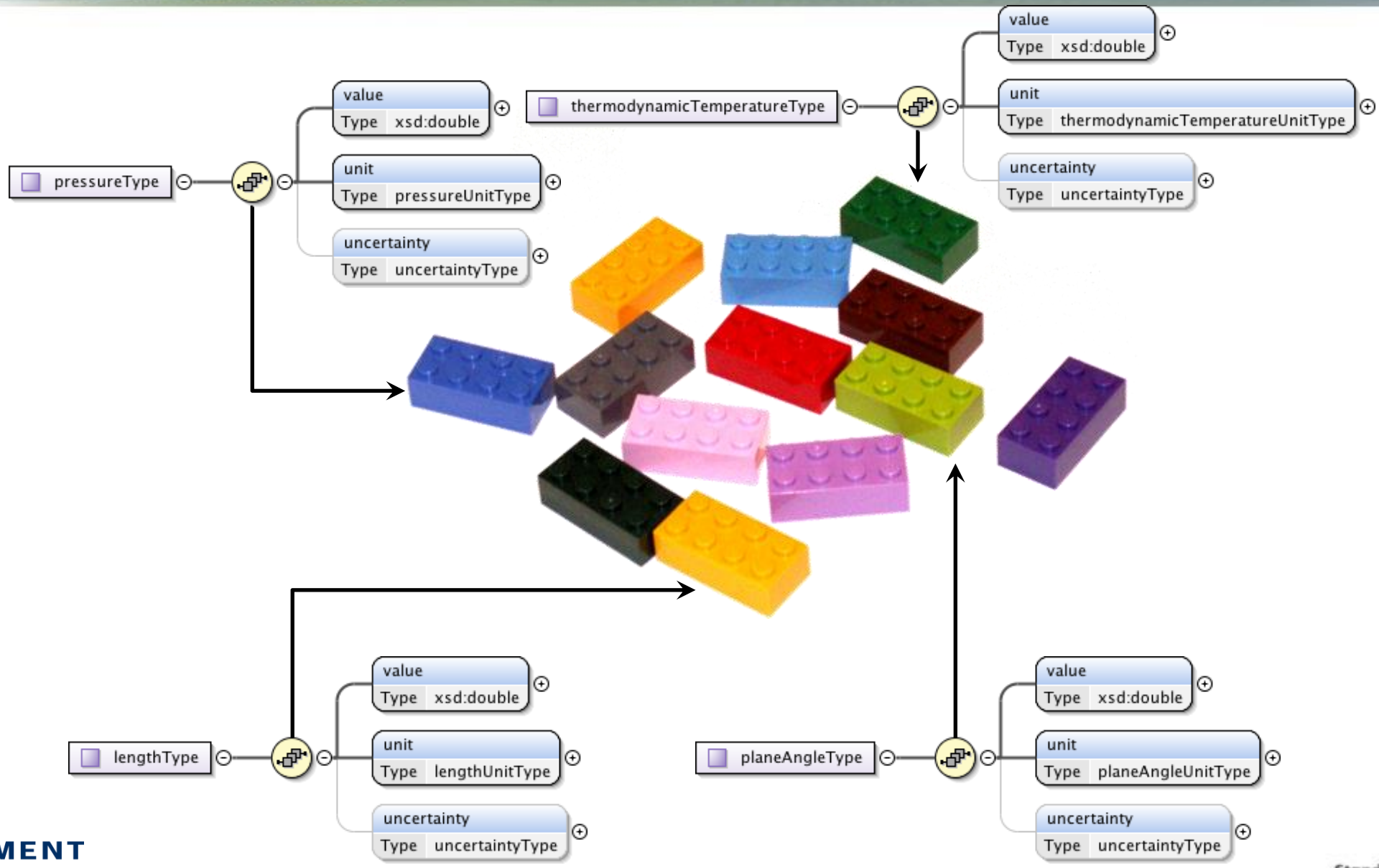
Images
Large Files
BLOBs

Database

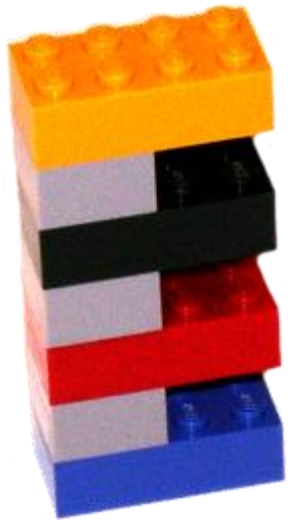
Large Dataset
Repository

Data Provider

Modularity: Foundational Types



Data Models Re-Use Components



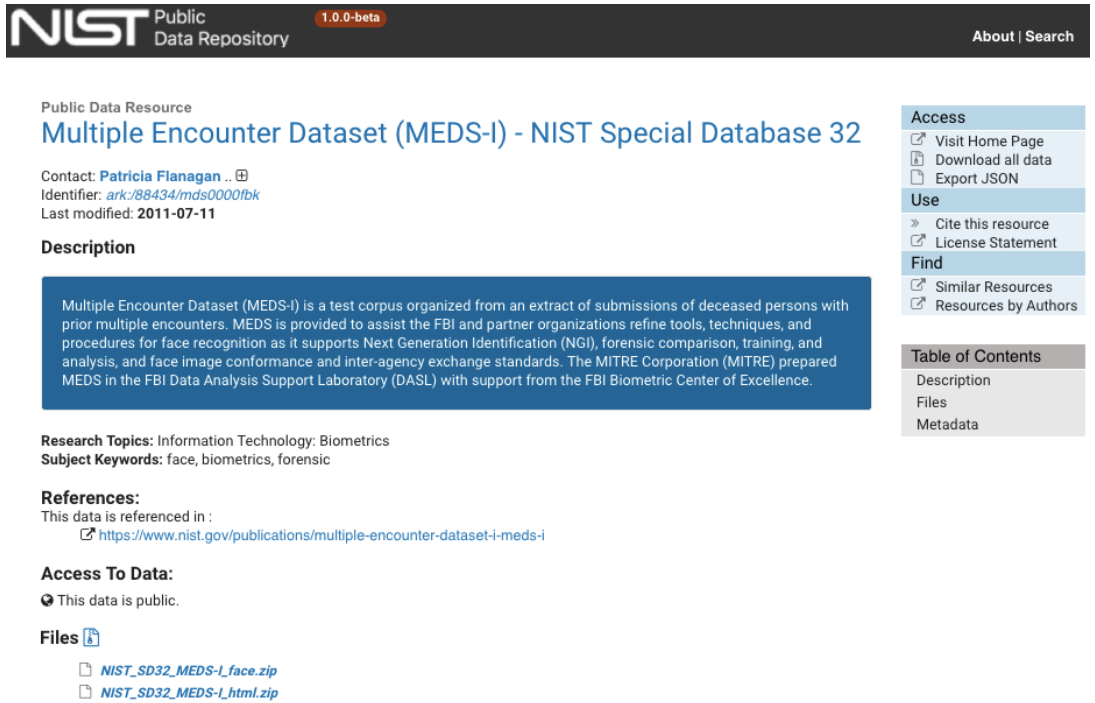
MATERIAL MEASUREMENT LABORATORY

NIST National Institute of Standards and Technology U.S. Department of Commerce

…and now over to Ray

NIST Public Data Repository

https://data.nist.gov/



Materials Resource Registry

https://materials.registry.nist.gov/

NIST National Institute of Standards and Technology U.S. Department of Commerce

PDR Preservation Service

- Public Data Repository specializes in “science-ready” data products
  - Standard Reference Data (SRD), Data Publications
  - Collections of files, ancillary data (previews, figures), metadata
- Preservation Service
  - package data for long-term storage
  - Standards-based
  - Packaging based on BagIt: NIST Preservation Profile

BagIt Packaging Standard

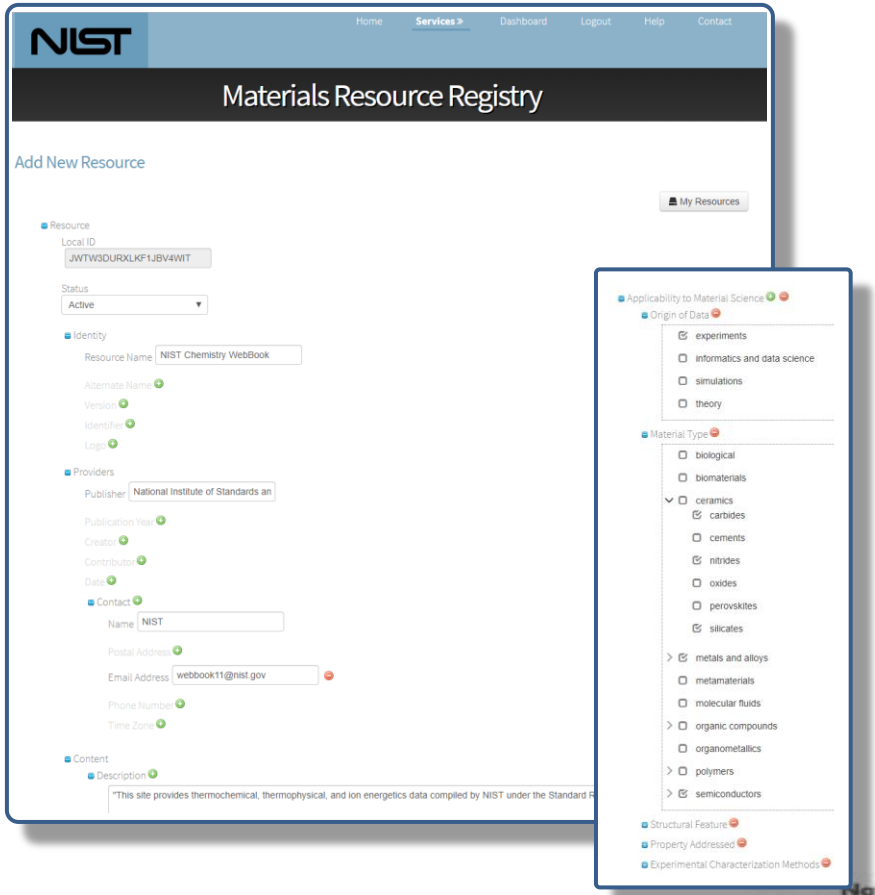
- Basics:
  - A Bag = a structured directory hierarchy (can be serialized, compressed)
  - **`data`** directory preserves native data organization
  - Manifest file provides checksums for each data file
  - Packaging metadata
  - Allows customization-–profiles—for storing metadata and other data
- Requirements: Transmission versus Preservation
  - Aggregates heterogeneous files, Information preserving
  - Self-describing, validate-able, corruption-detection
  - Scalability to large collections:
    - Transmission: **`fetch.txt`** lists URLs for externally retrieving files that are part of collection
    - Preservation: should not rely on existence of external service

Multibag BagIt Profile

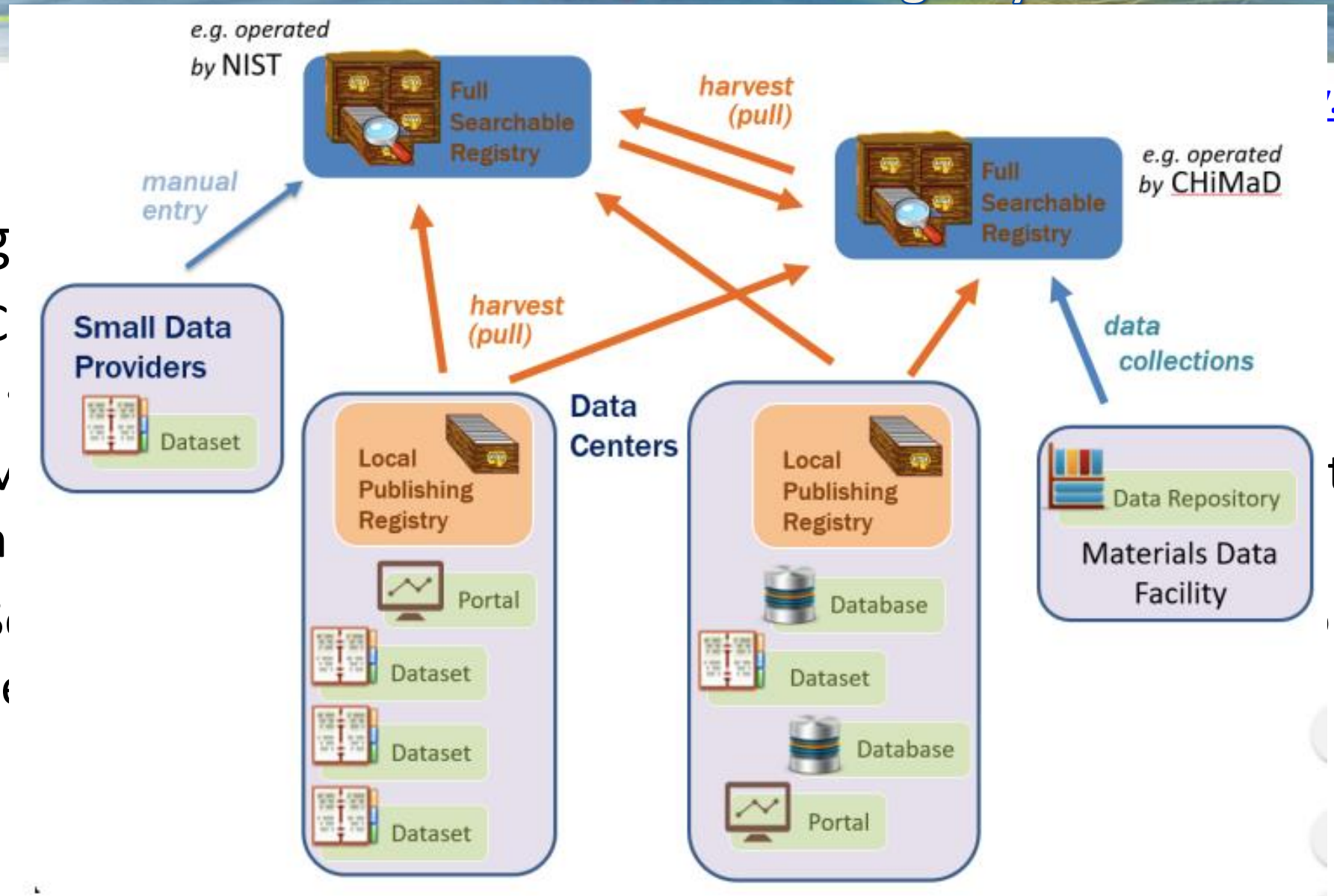
- Splits a logical bag over multiple component bags
- Drivers:
  - Data products that are large: total size or number of files
    - Allows for more efficient storage and retrieval
  - Efficient revisions to preserved data products
    - Create new component bags that contain only the parts that have changed
- Profile formalizes relationships between bags
  - “Head bag”: points to other component bags
  - Defines how to recombine components to reconstruct complete bag
  - Revisions create a new “head bag”

Materials Resource Registry:
Supporting Data Discovery

https://materials.registry.nist.gov/



Materials Resource Registry:



- Reg
  - C
  - M
    h
  - S
    se

.nist.gov/

tions via

n

Resource Metadata

https://materials.registry.nist.gov/

Metadata framework supports extensibility, evolution

- Generic core metadata + extensions
- Domain-specific extensions
  - Clients can ignore extensions it doesn't understand
- Materials Vocabulary
  - Modust but practical: 3 tiers of detail
  - See https://www.rd-alliance.org/materials-vocabulary-draft-21-mar-2017
  - SKOS version available
- Based on XML, XML Schema
  - Use of XSD Types to support plug-in extensions
- JSON-based version of techniques used in the PDR
  - JSON Schema: schema definition, types, validation
  - JSON-LD: tie definitions to community semantics